Text as Data

Justin Grimmer

Associate Professor
Department of Political Science
Stanford University

October 16th, 2014

Programming $\rightsquigarrow$ It Just Takes Time

Unigram Language Model

1) Task:
    - Summarize a document's content
    - Building block for many other models
2) Objective function $\rightsquigarrow$ Posterior distributions
    - Dirichlet-multinomial distribution
    - Logistic-normal distribution
3) Optimization
    - Maximum a Posteriori (MAP) values $\rightsquigarrow$ parameters that maximize the poserior
    - Use special properties
4) Validation
    - Overall: is our model distilling interesting information?

Building Models vs Algorithms

Today we're building statistical models

Building Models vs Algorithms

Today we're building statistical models
We'll still algorithms $\rightsquigarrow$ models without a data generating process

Building Models vs Algorithms

Today we're building statistical models
We'll still algorithms $\rightsquigarrow$ models without a data generating process
False debate. Both are useful

Building Models vs Algorithms

Today we're building statistical models

We'll still algorithms$\rightsquigarrow$ models without a data generating process

False debate. Both are useful

We will often use statistics, because:

Building Models vs Algorithms

Today we're building statistical models

We'll still algorithms $\rightsquigarrow$ models without a data generating process

False debate. Both are useful

We will often use statistics, because:

1) Clarity of assumptions

Building Models vs Algorithms

Today we're building statistical models
We'll still algorithms $\rightsquigarrow$ models without a data generating process
False debate. Both are useful
We will often use statistics, because:

1) Clarity of assumptions
2) Machinery to think about optimization, uncertainty, and sensitivity $\rightsquigarrow$ automatic objective function

Building Models vs Algorithms

Today we're building statistical models
We'll still algorithms $\rightsquigarrow$ models without a data generating process
False debate. Both are useful
We will often use statistics, because:

1) Clarity of assumptions
2) Machinery to think about optimization, uncertainty, and sensitivity $\rightsquigarrow$ automatic objective function
3) Easier to extend—leverage technology to generalize the model

Building Models vs Algorithms

Today we're building statistical models

We'll still algorithms $\rightsquigarrow$ models without a data generating process

False debate. Both are useful

We will often use statistics, because:

1) Clarity of assumptions
2) Machinery to think about optimization, uncertainty, and sensitivity $\rightsquigarrow$ automatic objective function
3) Easier to extend—leverage technology to generalize the model

Probability model $\rightsquigarrow$ form basis of statistical approaches

Unigram Model of Language

Assume we have a 3 word vocabulary

Unigram Model of Language

Assume we have a 3 word vocabulary $\rightsquigarrow$ 3 words that we might speak.

Unigram Model of Language

Assume we have a 3 word vocabulary $\rightsquigarrow$ 3 words that we might speak.
Bag of Words $\rightsquigarrow$ each word is an independent draw over 3 words

Unigram Model of Language

Assume we have a 3 word vocabulary $\rightsquigarrow$ 3 words that we might speak.
Bag of Words $\rightsquigarrow$ each word is an independent draw over 3 words

- Improbable model of language creation

Unigram Model of Language

Assume we have a 3 word vocabulary $\rightsquigarrow$ 3 words that we might speak.
Bag of Words $\rightsquigarrow$ each word is an independent draw over 3 words

- Improbable model of language creation
- Complex dependency structure of text

Unigram Model of Language

Assume we have a 3 word vocabulary$\rightsquigarrow$ 3 words that we might speak.
Bag of Words$\rightsquigarrow$ each word is an independent draw over 3 words

- Improbable model of language creation
- Complex dependency structure of text
- Improbable $\neq$ useless

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$p(\boldsymbol{X}_i = (1,0,0)) \quad = \quad \theta_1$$

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$
\begin{aligned}
p(\boldsymbol{X}_i = (1,0,0)) &= \theta_1 \\
p(\boldsymbol{X}_i = (0,1,0)) &= \theta_2
\end{aligned}
$$

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$\begin{aligned} p(\boldsymbol{X}_i = (1,0,0)) &= \theta_1 \\ p(\boldsymbol{X}_i = (0,1,0)) &= \theta_2 \\ p(\boldsymbol{X}_i = (0,0,1)) &= \theta_3 = 1 - \theta_2 - \theta_1 \end{aligned}$$

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$
\begin{aligned}
p(\boldsymbol{X}_i = (1,0,0)) &= \theta_1 \\
p(\boldsymbol{X}_i = (0,1,0)) &= \theta_2 \\
p(\boldsymbol{X}_i = (0,0,1)) &= \theta_3 = 1 - \theta_2 - \theta_1
\end{aligned}
$$

The pmf for $\boldsymbol{X}_i$ is,

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$\begin{aligned} p(\boldsymbol{X}_i = (1,0,0)) &= \theta_1 \\ p(\boldsymbol{X}_i = (0,1,0)) &= \theta_2 \\ p(\boldsymbol{X}_i = (0,0,1)) &= \theta_3 = 1 - \theta_2 - \theta_1 \end{aligned}$$

The pmf for $\boldsymbol{X}_i$ is,

$$p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \prod_{j=1}^{3} \theta_j^{x_{ij}}$$

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$\begin{aligned} p(\boldsymbol{X}_i = (1,0,0)) &= \theta_1 \\ p(\boldsymbol{X}_i = (0,1,0)) &= \theta_2 \\ p(\boldsymbol{X}_i = (0,0,1)) &= \theta_3 = 1 - \theta_2 - \theta_1 \end{aligned}$$

The pmf for $\boldsymbol{X}_i$ is,

$$\begin{aligned} p(\boldsymbol{x}_i|\boldsymbol{\theta}) &= \prod_{j=1}^{3} \theta_j^{x_{ij}} \\ \boldsymbol{X}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}) \end{aligned}$$

Unigram Model of Language

Suppose we are drawing a word $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})$

$$\begin{aligned} p(\boldsymbol{X}_i = (1,0,0)) &= \theta_1 \\ p(\boldsymbol{X}_i = (0,1,0)) &= \theta_2 \\ p(\boldsymbol{X}_i = (0,0,1)) &= \theta_3 = 1 - \theta_2 - \theta_1 \end{aligned}$$

The pmf for $\boldsymbol{X}_i$ is,

$$\begin{aligned} p(\boldsymbol{x}_i|\boldsymbol{\theta}) &= \prod_{j=1}^{3} \theta_j^{x_{ij}} \\ \boldsymbol{X}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}) \\ \boldsymbol{X}_i &\sim \text{Categorical}(\boldsymbol{\theta}) \end{aligned}$$

Unigram Model of Language

$$
\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\theta}) &= \prod_{j=1}^{3} \theta_j^{x_{ij}} \\
\boldsymbol{X}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta})
\end{aligned}
$$

Unigram Model of Language

$$p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \prod_{j=1}^{3} \theta_j^{x_{ij}}$$

$$\boldsymbol{X}_i \sim \text{Multinomial}(1, \boldsymbol{\theta})$$

$$E[x_{ij}] = \theta_j$$

Unigram Model of Language

$$\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\theta}) &= \prod_{j=1}^{3} \theta_j^{x_{ij}} \\
\boldsymbol{X}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}) \\
E[x_{ij}] &= \theta_j \\
\text{Var}(X_{ij}) &= \theta_j(1-\theta_j)
\end{aligned}$$

Unigram Model of Language

$$\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\theta}) &= \prod_{j=1}^{3} \theta_j^{x_{ij}} \\
\boldsymbol{X}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}) \\
E[x_{ij}] &= \theta_j \\
\text{Var}(X_{ij}) &= \theta_j(1-\theta_j) \\
\text{Cov}(X_{ij}, x_{ik}) &= -\theta_j\theta_k
\end{aligned}$$

Unigram Model of Language

Suppose we make N independent draws:

$$\boldsymbol{X}_i \sim \text{Multinomial}(1, \boldsymbol{\theta})$$

Then:

$$\begin{aligned}
\boldsymbol{X} &= \sum_{i=1}^{N} \boldsymbol{X}_i \\
&= (\sum_{i=1}^{N} X_{i1}, \sum_{i=1}^{N} X_{i2}, \sum_{i=1}^{N} X_{i3}) \\
\boldsymbol{X} &\sim \text{Multinomial}(N, \boldsymbol{\theta}) \\
p(\boldsymbol{x}|\boldsymbol{\theta}) &\propto \prod_{j=1}^{3} \theta_j^{x_j}
\end{aligned}$$

Unigram Model of Language

Obtaining maximum-likelihood estimates:

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) \propto \prod_{j=1}^{3} \theta_j^{x_j}$$

$$\log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{j=1}^{3} x_j \log \theta_j + c$$

Include constraint that $\sum_{j=1}^{3} \theta_j = 1$

$$\log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{j=1}^{3} x_j \log \theta_j + \lambda(\sum_{j=1}^{3} \theta_j - 1) + c$$

Unigram Model of Language

$$
\begin{aligned}
\log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) &= \sum_{j=1}^{3} x_j \log \theta_j + \lambda(\sum_{j=1}^{3} \theta_j - 1) + c \\
\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \theta_1} &= \frac{x_1}{\theta_1} + \lambda \\
\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \theta_2} &= \frac{x_2}{\theta_2} + \lambda \\
\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \theta_3} &= \frac{x_3}{\theta_3} + \lambda \\
\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \lambda} &= \sum_{j=1}^{3} \theta_j - 1
\end{aligned}
$$

Unigram Model of Language

$$
\begin{aligned}
0 &= \frac{x_1}{\theta_1^*} + \lambda^* \\
0 &= \frac{x_2}{\theta_2^*} + \lambda^* \\
0 &= \frac{x_3}{\theta_3^*} + \lambda^* \\
0 &= \sum_{j=1}^{3} \theta_j^* - 1
\end{aligned}
$$

Unigram Model of Language

$$\theta_1^* = \frac{x_1}{x_1 + x_2 + x_3}$$
$$\theta_2^* = \frac{x_2}{x_1 + x_2 + x_3}$$
$$\theta_3^* = \frac{x_3}{x_1 + x_2 + x_3}$$

Unigram Model of Language

$$\theta_1^* = \frac{x_1}{x_1 + x_2 + x_3}$$
$$\theta_2^* = \frac{x_2}{x_1 + x_2 + x_3}$$
$$\theta_3^* = \frac{x_3}{x_1 + x_2 + x_3}$$

Maximum likelihood estimates $\rightsquigarrow$ Rates words are used

Unigram Model of Language

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^{3} \theta_j^{x_j}$$

Unigram Model of Language

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^{3} \theta_j^{x_j}$$

$\boldsymbol{\theta}$: encodes information about word rates $\rightsquigarrow$ our summary of the document/speaker

Unigram Model of Language

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^{3} \theta_j^{x_j}$$

$\boldsymbol{\theta}$: encodes information about word rates$\rightsquigarrow$ our summary of the document/speaker

- $\sum_{j=1}^{3} \theta_j = 1$

Unigram Model of Language

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \quad \propto \quad \prod_{j=1}^{3} \theta_j^{x_j}$$

$\boldsymbol{\theta}$: encodes information about word rates$\rightsquigarrow$ our summary of the document/speaker

- $\sum_{j=1}^{3} \theta_j = 1$
- $\theta_j \geq 0$

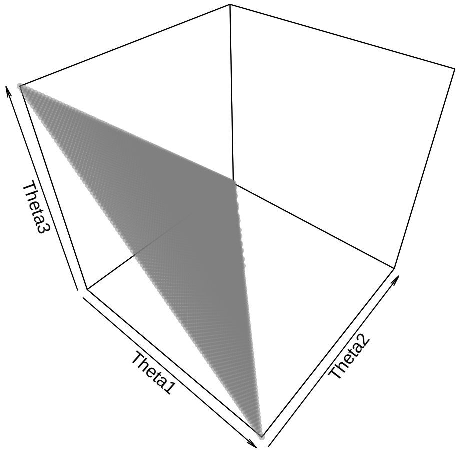
Unigram Model of Language

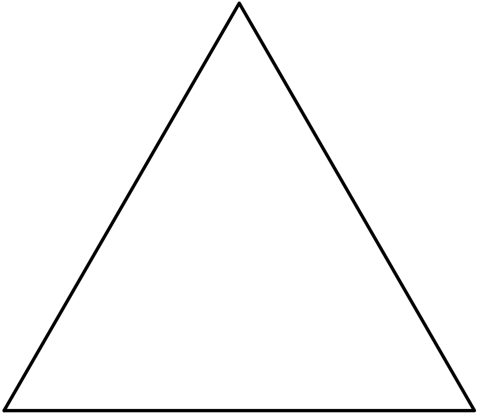
$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto \prod_{j=1}^{3} \theta_j^{x_j}$$

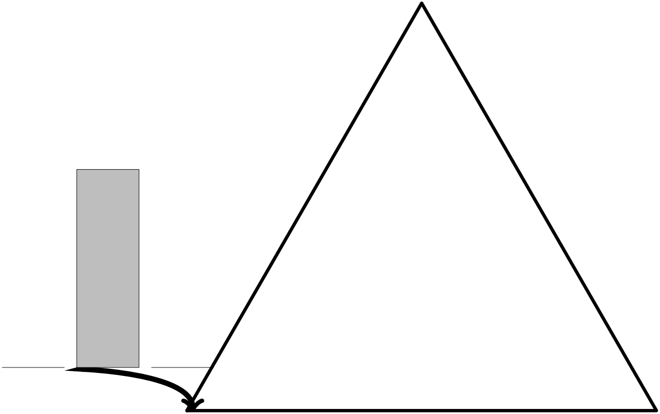
$\boldsymbol{\theta}$: encodes information about word rates$\rightsquigarrow$ our summary of the document/speaker

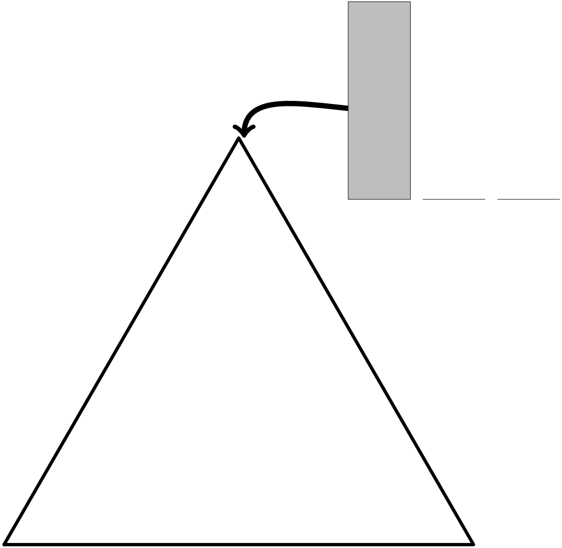
- $\sum_{j=1}^{3} \theta_j = 1$
- $\theta_j \geq 0$

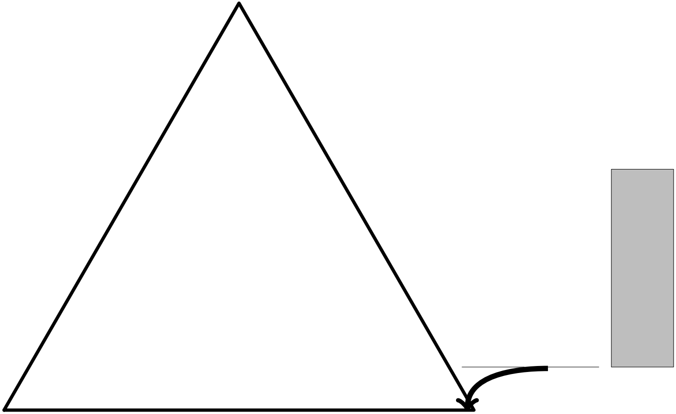
$\boldsymbol{\theta} \in \Delta^2$ (2-dimensional simplex)

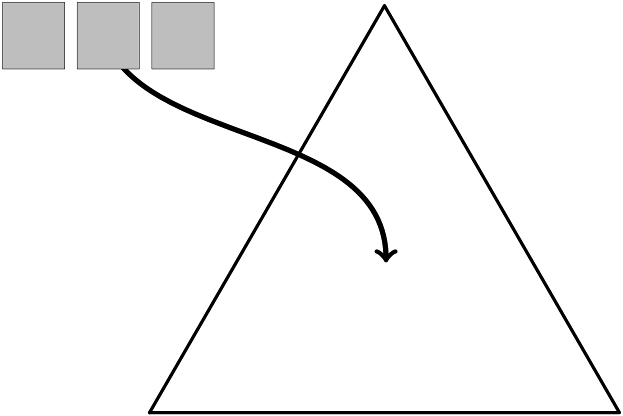

Theta3
Theta1
Theta2











Unigram Model of Language

Suppose we have several speakers (authors/clusters/topics/categories/ ...)
Speaker i produces document $\boldsymbol{x}_i$,

$$\boldsymbol{X}_i \sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i \rightsquigarrow$ Speaker specific word rates
Build hierarchical model:

$$\boldsymbol{\theta}_i \sim \text{Distribution on Simplex}$$

Hierarchical Models as a Modeling Paradigm

Why Build a Hierarchical Model?

1) Borrow strength across documents$\rightsquigarrow$ Improved and granular inferences
2) Shrink estimates$\rightsquigarrow$ regularization
3) Incorporate further covariate information
   i) Author
   ii) Time
   iii) ...
3) Learn additional structure
   i) Hierarchies of word rates
   ii) Clusters of similar word rates
   iii) Low dimensional approximations of word rates
4) Encodes complicated dependencies between documents/speakers

Dirichlet-Multinomial Unigram Language Model

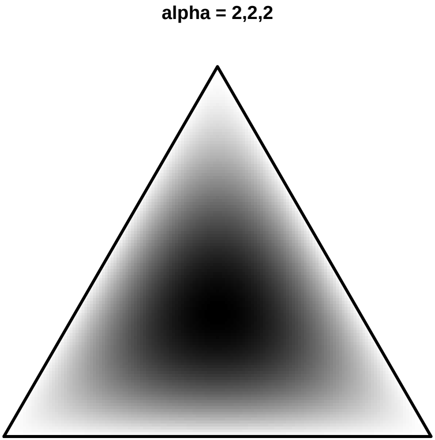
For N observations we observe a 3-element long count vector

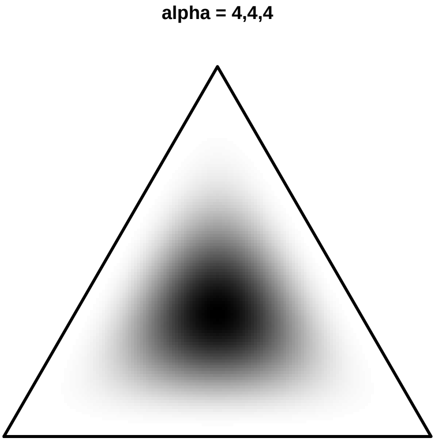
$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})$$

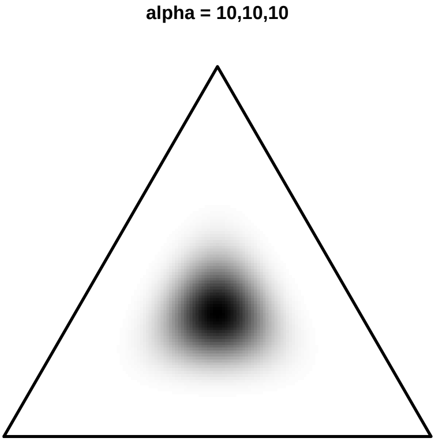
Where $N_i = \sum_{j=1}^{3} x_{ij}$.
Suppose

$$\begin{aligned} \boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i) \end{aligned}$$

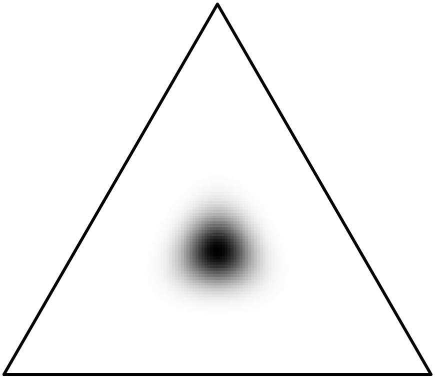
- Dirichlet distribution $\rightsquigarrow$ assumption about population of word rates
- $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ describes population use of words and variation
- Just one distribution simplex


alpha = 2,2,2

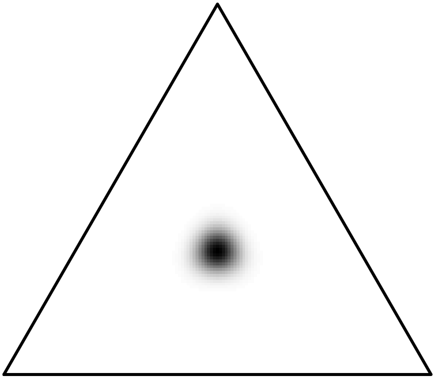

alpha = 4,4,4

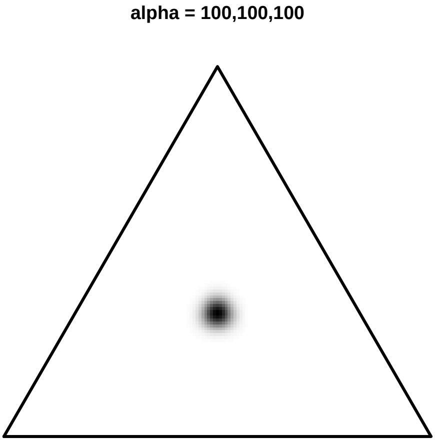

alpha = 10,10,10

alpha = 20,20,20

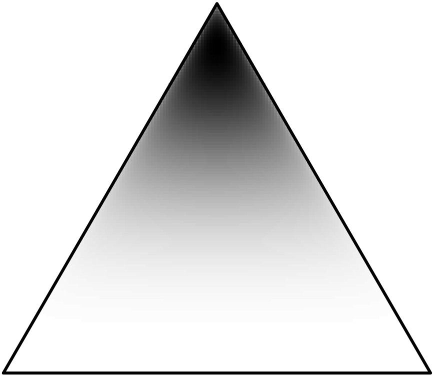


alpha = 50,50,50

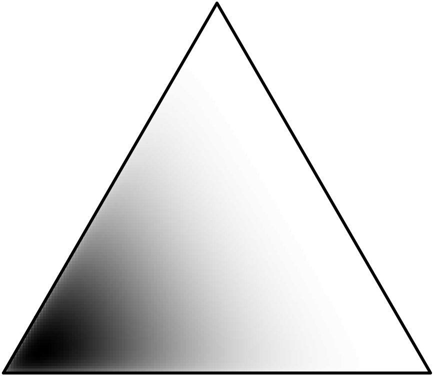



alpha = 100,100,100

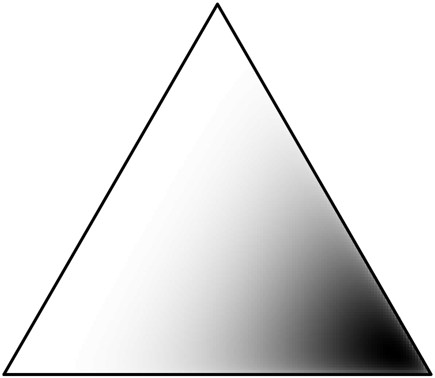
alpha = 4,1.2,1.2



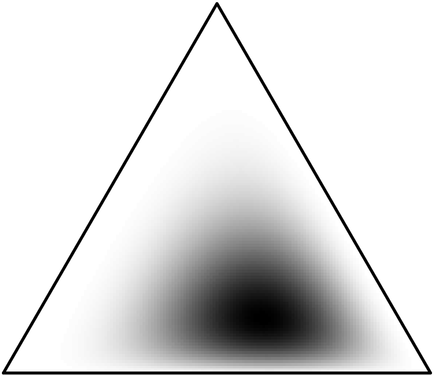
alpha = 1.2,4,1.2



alpha = 1.2,1.2,4



alpha = 2.04,3.24,4.72



Dirichlet Distribution

Suppose

$$\boldsymbol{\theta}_i \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

Then,

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{3} \alpha_j)}{\prod_{j=1}^{3} \Gamma(\alpha_j)} \prod_{j=1}^{3} \theta_{ij}^{\alpha_j - 1}$$

- If $\boldsymbol{\alpha} = (1, 1, 1)$ Uniform distribution

Dirichlet Distribution

- Important Facts

$$
\begin{aligned}
E[\boldsymbol{\theta}_i] &= \left(\frac{\alpha_1}{\sum_{j=1}^{3}\alpha_j}, \frac{\alpha_2}{\sum_{j=1}^{3}\alpha_j}, \frac{\alpha_3}{\sum_{j=1}^{3}\alpha_j}\right) \\
\text{var}(\theta_{ij}) &= \frac{\alpha_i\left(\sum_{j=1}^{3}\alpha_j - \alpha_i\right)}{\left(\sum_{j=1}^{3}\alpha_j\right)^2\left(\sum_{j=1}^{3}\alpha_j + 1\right)} \\
\text{cov}(\theta_{ik}, \theta_{ij}) &= \frac{-\alpha_k\alpha_j}{\left(\sum_{j=1}^{3}\alpha_j\right)^2\left(\sum_{j=1}^{3}\alpha_j + 1\right)} \\
\text{Mode}(\theta_j) &= \frac{\alpha_j - 1}{\sum_{k=1}^{3}\alpha_k - 3}
\end{aligned}
$$

Dirichlet-Multinomial Unigram Model of Language

$$\begin{aligned} \boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i) \end{aligned}$$

Dirichlet-Multinomial Unigram Model of Language

$$
\begin{aligned}
\boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_i | \boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)
\end{aligned}
$$

$$
p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}, \boldsymbol{x}_i) \propto p(\boldsymbol{\theta}_i | \boldsymbol{\alpha})\ p(\boldsymbol{x}_i | \boldsymbol{\theta}_i)
$$

Dirichlet-Multinomial Unigram Model of Language

$$
\begin{aligned}
\boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}, \boldsymbol{x}_i) &\propto p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}) \; p(\boldsymbol{x}_i|\boldsymbol{\theta}_i) \\
&\propto \frac{\Gamma(\sum_{j=1}^{3} \alpha_j)}{\prod_{j=1}^{3} \Gamma(\alpha_j)} \prod_{j=1}^{3} \theta_j^{\alpha_j - 1} \prod_{j=1}^{3} \theta_{ij}^{x_{ij}}
\end{aligned}
$$

Dirichlet-Multinomial Unigram Model of Language

$$\begin{aligned}\boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)\end{aligned}$$

$$\begin{aligned}p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}, \boldsymbol{x}_i) &\propto p(\boldsymbol{\theta}_i|\boldsymbol{\alpha})\ p(\boldsymbol{x}_i|\boldsymbol{\theta}_i) \\ &\propto \frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\prod_{j=1}^{3}\Gamma(\alpha_j)}\prod_{j=1}^{3}\theta_j^{\alpha_j-1}\prod_{j=1}^{3}\theta_{ij}^{x_{ij}} \\ &\propto \frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\prod_{j=1}^{3}\Gamma(\alpha_j)}\underbrace{\prod_{j=1}^{3}\theta_j^{\alpha_j+x_{ij}-1}}_{\text{Dirichlet Kernel}}\end{aligned}$$

Dirichlet-Multinomial Unigram Model of Language

$$\begin{aligned} \boldsymbol{\theta}_i | \boldsymbol{\alpha}, \boldsymbol{x}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha} + \boldsymbol{x}) \\ \text{E}[\theta_{ij} | \boldsymbol{\alpha}, \boldsymbol{x}_i] &= \frac{\alpha_j + x_{ij}}{\sum_{j=1}^{3} (x_{ij} + \alpha_j)} \end{aligned}$$

- $\alpha_j \rightsquigarrow$ "pseudo" data that smooth the estimates toward $\frac{\alpha_j}{\alpha_1 + \alpha_2 + \alpha_3}$
- as $N_i \to \infty$ data ($\boldsymbol{x}_i$) overwhelm $\boldsymbol{\alpha}$

Dirichlet-Multinomial Unigram Model of Language

Data generation process suggests new probability mass function for $\boldsymbol{x}_i \rightsquigarrow$ marginalize over $\boldsymbol{\theta}$

$$
\begin{aligned}
\boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)
\end{aligned}
$$

Dirichlet-Multinomial Unigram Model of Language

Data generation process suggests new probability mass function for $\boldsymbol{x}_i \rightsquigarrow$ marginalize over $\boldsymbol{\theta}$

$$
\begin{aligned}
\boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)
\end{aligned}
$$

This implies:

Dirichlet-Multinomial Unigram Model of Language

Data generation process suggests new probability mass function for $\boldsymbol{x}_i \rightsquigarrow$ marginalize over $\boldsymbol{\theta}$

$$\begin{aligned} \boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_i | \boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i) \end{aligned}$$

This implies:

$$p(\boldsymbol{x}_i | \boldsymbol{\alpha}) = \int_{\Delta^2} p(\boldsymbol{x}_i, \boldsymbol{\theta}_i | \boldsymbol{\alpha}) d\boldsymbol{\theta}$$

Dirichlet-Multinomial Unigram Model of Language

Data generation process suggests new probability mass function for $\boldsymbol{x}_i \rightsquigarrow$ marginalize over $\boldsymbol{\theta}$

$$\begin{aligned} \boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i) \end{aligned}$$

This implies:

$$\begin{aligned} p(\boldsymbol{x}_i|\boldsymbol{\alpha}) &= \int_{\Delta^2} p(\boldsymbol{x}_i, \boldsymbol{\theta}_i|\boldsymbol{\alpha}) d\boldsymbol{\theta} \\ &= \int_{\Delta^2} p(\boldsymbol{x}_i|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}) d\boldsymbol{\theta} \end{aligned}$$

Dirichlet-Multinomial Unigram Model of Language

Data generation process suggests new probability mass function for $\boldsymbol{x}_i \rightsquigarrow$ marginalize over $\boldsymbol{\theta}$

$$
\begin{aligned}
\boldsymbol{\theta}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{x}_i|\boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)
\end{aligned}
$$

This implies:

$$
\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\alpha}) &= \int_{\Delta^2} p(\boldsymbol{x}_i, \boldsymbol{\theta}_i|\boldsymbol{\alpha}) d\boldsymbol{\theta} \\
&= \int_{\Delta^2} p(\boldsymbol{x}_i|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}) d\boldsymbol{\theta} \\
&= \binom{N_i}{n_1! n_2! n_3!} \frac{\Gamma(\sum_{j=1}^{3} \alpha_j)}{\prod_{j=1}^{3} \Gamma(\alpha_j)} \int_{\Delta^2} \prod_{j=1}^{3} \theta_{ij}^{x_{ij}} \theta_{ij}^{\alpha_j - 1} d\boldsymbol{\theta}
\end{aligned}
$$

$$p(\boldsymbol{x}_i|\boldsymbol{\alpha}) = \binom{N_i}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\prod_{j=1}^{3}\Gamma(\alpha_j)}\int_{\Delta^2}\underbrace{\prod_{j=1}^{3}\theta_{ij}^{x_{ij}}\theta_{ij}^{\alpha_j-1}d\boldsymbol{\theta}}_{\text{Dirichlet Kernel}}$$

$$
\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\alpha}) &= \binom{N_i}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\prod_{j=1}^{3}\Gamma(\alpha_j)}\int_{\Delta^2}\underbrace{\prod_{j=1}^{3}\theta_{ij}^{x_{ij}}\theta_{ij}^{\alpha_j-1}d\boldsymbol{\theta}}_{\text{Dirichlet Kernel}} \\
&= \binom{N_i}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\prod_{j=1}^{3}\Gamma(\alpha_j)}\frac{\prod_{j=1}^{3}\Gamma(x_{ij}+\alpha_j)}{\Gamma(\sum_{j=1}^{3}(x_{ij}+\alpha_j))}
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\alpha}) &= \binom{N_i}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\prod_{j=1}^{3}\Gamma(\alpha_j)}\int_{\Delta^2}\underbrace{\prod_{j=1}^{3}\theta_{ij}^{x_{ij}}\theta_{ij}^{\alpha_j-1}d\boldsymbol{\theta}}_{\text{Dirichlet Kernel}} \\
&= \binom{N_i}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\prod_{j=1}^{3}\Gamma(\alpha_j)}\frac{\prod_{j=1}^{3}\Gamma(x_{ij}+\alpha_j)}{\Gamma(\sum_{j=1}^{3}(x_{ij}+\alpha_j))} \\
&= \binom{N_i}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\Gamma(\sum_{j=1}^{3}(x_{ij}+\alpha_j))}\prod_{j=1}^{3}\frac{\Gamma(x_{ij}+\alpha_j)}{\Gamma(\alpha_j)}
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\alpha}) &= \binom{N_i}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\prod_{j=1}^{3}\Gamma(\alpha_j)}\int_{\Delta^2}\underbrace{\prod_{j=1}^{3}\theta_{ij}^{x_{ij}}\theta_{ij}^{\alpha_j-1}d\boldsymbol{\theta}}_{\text{Dirichlet Kernel}} \\
&= \binom{N_i}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\prod_{j=1}^{3}\Gamma(\alpha_j)}\frac{\prod_{j=1}^{3}\Gamma(x_{ij}+\alpha_j)}{\Gamma(\sum_{j=1}^{3}(x_{ij}+\alpha_j))} \\
&= \binom{N_i}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\Gamma(\sum_{j=1}^{3}(x_{ij}+\alpha_j))}\prod_{j=1}^{3}\frac{\Gamma(x_{ij}+\alpha_j)}{\Gamma(\alpha_j)}
\end{aligned}
$$

Has some intuitive properties

$$
\begin{aligned}
p(\boldsymbol{x}_i|\boldsymbol{\alpha}) &= \binom{N_i}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)}\int_{\Delta^2}\underbrace{\prod_{j=1}^{3}\theta_{ij}^{x_{ij}}\theta_{ij}^{\alpha_j-1}d\boldsymbol{\theta}}_{\text{Dirichlet Kernel}} \\
&= \binom{N_i}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\prod_{j=1}^3 \Gamma(\alpha_j)}\frac{\prod_{j=1}^3 \Gamma(x_{ij}+\alpha_j)}{\Gamma(\sum_{j=1}^3(x_{ij}+\alpha_j))} \\
&= \binom{N_i}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^3 \alpha_j)}{\Gamma(\sum_{j=1}^3(x_{ij}+\alpha_j))}\prod_{j=1}^{3}\frac{\Gamma(x_{ij}+\alpha_j)}{\Gamma(\alpha_j)}
\end{aligned}
$$

Has some intuitive properties

$$
\mathrm{E}[X_{ij}] = N\frac{\alpha_j}{\sum_{k=1}^3 \alpha_k}
$$

Dirichlet-Multinomial Unigram Model

We can also generate a predictive distribution $\rightsquigarrow$ probability next word is j

$$
\begin{aligned}
P(\tilde{x} = 1|\boldsymbol{x}_i, \boldsymbol{\alpha}) &= \int_{\Delta^2} p(\tilde{x} = 1|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{x}_i) d\boldsymbol{\theta} \\
&= \int_{\Delta^2} \theta_j \text{Dir}(\boldsymbol{\theta}|\boldsymbol{x}_i + \boldsymbol{\alpha}) d\boldsymbol{\theta} \\
&= \frac{x_{ij} + \alpha_j}{\sum_{j=1}^{3} (x_{ij} + \alpha_j)}
\end{aligned}
$$

Dirichlet-Multinomial Unigram Model

Where does $\boldsymbol{\alpha}$ come from?
Extend the model $\rightsquigarrow$ infer it

$$
\begin{aligned}
\alpha_j &\sim \text{Gamma}(0.25, 1) \\
\boldsymbol{\theta}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\theta}_i) \\
\boldsymbol{x}_i | \boldsymbol{\theta}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)
\end{aligned}
$$

Dirichlet-Multinomial Unigram Model

Which yields

$$
\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{\alpha} | \boldsymbol{X}) &\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) p(\boldsymbol{x}_i | \boldsymbol{\alpha}) \\
p(\boldsymbol{\alpha} | \boldsymbol{X}) &= \int_{\Delta^2} p(\boldsymbol{\theta}, \boldsymbol{\alpha} | \boldsymbol{X}) d\boldsymbol{\theta} \\
&\propto p(\boldsymbol{\alpha}) \prod_{i=1}^{N} \int_{\Delta^2} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) p(\boldsymbol{x}_i | \boldsymbol{\alpha}) d\boldsymbol{\theta} \\
&\propto \prod_{j=1}^{3} 4 \exp(-4\alpha_j) \times \prod_{i=1}^{N} \left[\frac{\Gamma(\sum_{j=1}^{3} \alpha_j)}{\Gamma(\sum_{j=1}^{3} (x_{ij} + \alpha_j))} \times \prod_{j=1}^{3} \frac{\Gamma(x_{ij} + \alpha_j)}{\Gamma(\alpha_j)} \right]
\end{aligned}
$$

Unigram Model of Language

Suppose $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$

- $x_{ij} =$ Number of times word j occurs in document i.
- $N_i = \sum_{j=1}^{J} x_{ij}$ total number of words in document i

Assume a generation process

$$
\begin{aligned}
\boldsymbol{x}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}) \\
\boldsymbol{\theta} &= (\theta_1, \theta_2, \ldots, \theta_J) \\
p(\boldsymbol{x}_i | \boldsymbol{\theta}) &\propto \prod_{j=1}^{J} \theta_j^{x_{ij}}
\end{aligned}
$$

Alternative Priors on the Simplex

Dirichlet distribution

- Imposes specific form on variance
- Imposes negative correlation between all components.
- We might expect some word rates to positively covary.

Alternative $\rightsquigarrow$ Logistic-Normal distribution

Logistic-Normal Distribution

Suppose $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})$.
Define:

$$
\begin{aligned}
\boldsymbol{y}_i &= \left(\log\left(\frac{x_{i1}}{x_{i3}}\right), \log\left(\frac{x_{i2}}{x_{i3}}\right)\right) \\
\boldsymbol{y}_i &\sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\mu} &= (\mu_1, \mu_2) \\
\boldsymbol{\Sigma} &= \begin{pmatrix} \sigma_1^2 & \text{cov}(y_{i1}, y_{i2}) \\ \text{cov}(y_{i1}, y_{i2}) & \sigma_2^2 \end{pmatrix}
\end{aligned}
$$

Logistic-Normal Distribution

To get back original data apply:

$$
\begin{aligned}
x_{i1} &= \left(\frac{\exp(y_{i1})}{\exp(y_{i1}) + \exp(y_{i2}) + 1}\right) \\
x_{i2} &= \left(\frac{\exp(y_{i2})}{\exp(y_{i1}) + \exp(y_{i2}) + 1}\right) \\
x_{i3} &= \left(\frac{1}{\exp(y_{i1}) + \exp(y_{i2}) + 1}\right) \\
\boldsymbol{x}_i &= g(\boldsymbol{y}_i)
\end{aligned}
$$

Logistic-Normal Distribution

An alternative hierarchical model:

$$
\begin{aligned}
\boldsymbol{\eta}_i &\sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\theta}_i &= g(\boldsymbol{\eta}_i) \\
\boldsymbol{x}_i &\sim \text{Multinomial}(N_i, \boldsymbol{\theta}_i)
\end{aligned}
$$

Widely used:

- Correlated models
- Natural way to encode regressions in prior

Next week: clustering!